Compensating inaccurate annotations to train 3D facial landmark localization models

Federico M. Sukno*†, John L. Waddington† and Paul F. Whelan*
*Centre for Image Processing & Analysis, Dublin City University, Dublin 9, Ireland
†Molecular & Cellular Therapeutics, Royal College of Surgeons in Ireland, Dublin 2, Ireland

***Abstract*—In this paper we investigate the impact of inconsistency in manual annotations when they are used to train automatic models for 3D facial landmark localization. We start by showing that it is possible to objectively measure the consistency of annotations in a database, provided that it contains replicates (i.e. repeated scans from the same person). Applying such measure to the widely used FRGC database we find that manual annotations currently available are suboptimal and can strongly impair the accuracy of automatic models learnt therefrom. To address this issue, we present a simple algorithm to automatically correct a set of annotations and show that it can help to significantly improve the accuracy of the models in terms of landmark localization errors. This improvement is observed even when errors are measured with respect to the original (not corrected) annotations. However, we also show that if errors are computed against an alternative set of manual annotations with higher consistency, the accuracy of the models constructed using the corrections from the presented algorithm tends to converge to the one achieved by building the models on the alternative, more consistent set.**

I. INTRODUCTION

The localization of prominent facial landmarks is important for the majority of 3D facial analysis systems. In the context of facial biometrics, landmarks can be used either as the primary source of information [7] or merely as a detection and/or normalization step [14], but in both cases the accuracy of localization is an important factor that can condition the final performance of the whole system.

Thus, localization of facial landmarks in 3D can be considered a relevant topic in itself and has attracted considerable attention, including the deployment of annotated datasets to benchmark different algorithms. For example, there are the publicly available annotations for nearly 5000 scans from the Face Recognition Grand Challenge (FRGC) database [15], which constitutes one of the most widely used datasets to report localization accuracy of 3D facial landmarks.

The results from automatic approaches indicate that the most prominent facial landmarks can be located with errors varying between 3 and 6 mm [14], [19], [22], [25], [26], with some advantage to algorithms incorporating texture over those based purely on geometric features. However, these errors seem still far from the localization accuracy that might be achieved by means of manual annotations. Indeed, results from clinical research suggest that the errors of manual annotations for several facial landmarks can be as low as 1 to 2 mm [8], [16], [23].

We will show in Section II that the above discrepancy is partly due to the lack of consistency of the manual annotations currently available for FRGC. In contrast to traditional measures of accuracy, such as inter- and intra-observer variability, we base our analysis on the consistency of annotations by comparing the inter-landmark distances of replicates (i.e. different scans from the same individual). It is widely accepted that, except for the lower part of the face (mouth and chin), the pairwise distances between anthropometric landmarks should remain unchanged for different scans of the same individual. Thus, we can objectively measure how consistent are the annotations on a given dataset without the need to generate repeated markups.

Notice that consistency of annotations is a necessary but not sufficient condition for accuracy. Hence, lack of consistency implies lack of accuracy, with negative effects not only on the evaluation results but also on the accuracy of any model that is created using these annotations as a training set. The latter relates to the problem of learning with *noisy data*, which has been extensively studied in machine learning [9], [11], [13], [24]. The problem of inaccurate annotations can be thought of as class-label noise (i.e. the wrong coordinates in the facial scan are labeled as the ground truth landmark position), as opposed to attribute noise which occurs when the uncertainty affects primarily the extracted features (e.g. acquisition noise).

It has been shown that the impact of class-label noise in learning algorithms is twofold: 1) it reduces the classification accuracy, and 2) it increases the complexity of the classifier (when this is allowed by the algorithm, e.g. if using support vector machines or decision trees [24]). A popular approach to mitigate these effects has been trying to identify (and eliminate) the samples that are mislabeled. Examples of this strategy include the use of classifier ensembles to confirm the proposed labels by majority voting or consensus [1], [24], minimization of model's complexity [6] (under the assumption that eliminating mislabeled examples will reduce the complexity of the correct hypothesis), removing examples with poor stability of their labels based on a leave-one-out perturbation matrix [12] or with low probability of correct classification based on their neighbors [18].

An interesting difference in our case is that for each mislabeled sample we certainly know that there is a correct sample readily available. That is, a set of coordinates incorrectly labeled as the ground truth position of the nose tip could be ideally replaced by the *correct* set of coordinates, which are hopefully not too far away. Thus, we do not need to discard these samples but we may actually attempt to correct them. With this in mind, we present in Section III an algorithm that aims to automatically correct the annotations on a training set. We work under the hypotheses that the majority of annotations

are approximately correct and that a local geometry descriptor can be used to estimate corresponding points across different surfaces. The corrected annotations are obtained as those with the Least Squared Corrections of Uncertainty (LSCU) from the initial ones that achieve maximum similarity of the local geometry descriptor for a given *uncertainty radius*. This radius is the only parameter of the algorithm and indicates the maximum noise level that we expect from the input annotations.

We validate the correction algorithm in Section IV in the context of automatic landmark localization. For this purpose we select the SRILF algorithm (Shape Regression with Incomplete Local Features [20]), which learns both local descriptors and global shape statistics from a training set and has been shown to achieve overall errors of approximately 3 mm on clinical datasets. To test the impact of the corrected annotations, the algorithm is first trained with the original annotations and later trained with corrected annotations for various radii of uncertainty. Results demonstrate that models trained on data corrected by LSCU clearly outperform the models trained on the original (inconsistent) data, even when errors are computed using the latter as ground truth. We also compare the accuracy of the models with respect to a second set of *cleaner* annotations (in terms of consistency) and show that, if the annotations from the *noisy set* are corrected by LSCU, we can construct automatic models with similar accuracy to those built from the *cleaner* annotations.

II. Evaluating the consistency of annotations

Given a set of facial scans with landmark annotations (either manual or automatic), if the set has more than one scan per person (replicates) we can use the invariance of distances of certain facial landmarks to estimate the consistency of annotations.

Let $\{\mathbf{a}_i\}_{i=1}^N$ be a set of 3D annotations for N facial scans containing L landmarks each:

$$\mathbf{a}_i = (x_{i,1}, y_{i,1}, z_{i,1}, \ldots, x_{i,L}, y_{i,L}, z_{i,L})^T \tag{1}$$

and let $\mathrm{id}(\mathbf{a}_i)$ be the *identity* of the facial scan associated with $\mathbf{a}_i$. Then, for any pair of landmarks (ℓ_p, ℓ_q) for which their distance can be considered invariant to factors other than identity (e.g. expressions) the following should hold:

$$d\big(\mathbf{a}_i(\ell_p) - \mathbf{a}_i(\ell_q)\big) = d\big(\mathbf{a}_j(\ell_p) - \mathbf{a}_j(\ell_q)\big) \quad \forall (i,j) \,|\, \mathrm{id}(\mathbf{a}_i) = \mathrm{id}(\mathbf{a}_j) \tag{2}$$

where $\mathbf{a}_i(\ell_p) = (x_{i,p}, y_{i,p}, z_{i,p})^T$ are the coordinates of the *p-th* landmark and $d()$ is the Euclidean distance. Note that the above equality holds because 3D scanners provide World coordinates. To make a similar comparison in 2D we would need to either know the calibration matrix of the camera or use projective invariants, such as distance ratios.

The FRGC database provides a large collection of 3D scans with abundant replicates. For the experiments in this paper we manually annotated 11 landmarks[1] on the first 100 scans from FRGC (v1) and compared their consistency against the publicly available annotations from Szeptycki et al. [22], with some additions and corrections introduced by Creusot et al. [4][2]. This set contains scans from 19 different persons and allows for a total of 248 pairwise comparisons[3]. We measure the discrepancy of pairwise distances from the idealized case indicated in eq. (2):

$$e_{p,q}^{PWD} = \Big\{ \big| d\big(\mathbf{a}_i(\ell_p) - \mathbf{a}_i(\ell_q)\big) - d\big(\mathbf{a}_j(\ell_p) - \mathbf{a}_j(\ell_q)\big) \big| \Big\} \quad \forall (i,j) \,|\, \mathrm{id}(\mathbf{a}_i) = \mathrm{id}(\mathbf{a}_j) \tag{3}$$

That is, we measure the discrepancy of the distance between landmarks p and q measured from different scans of the same person. Fig. 1 shows the average discrepancy over the 248 pairwise comparisons for all possible landmark combinations, using both the publicly available ground-truth annotations (GTA-1) and our own manual annotations (GTA-2). The difference is substantial and not only restricted to the average, which we illustrate by displaying boxplots of some landmark pairs in Fig. 2. It is interesting to note that, when we consider landmarks that do vary under expression changes, such as mouth corners (ch), the difference between both sets of annotations (in terms of consistency) are considerably reduced. In these cases we can have values of $e_{p,q}^{PWD} > 0$ due to inconsistencies of the annotations and/or due to expression changes.

However, when considering landmarks from the nose and the eyes, inter-landmark distances are almost invariant for replicate scans and there are large and statistically significant differences between both sets of annotations. This is easily confirmed by visual inspection, as illustrated in Fig. 3; it is clear that the criteria used to annotate landmark positions have not been applied homogeneously across the database. This hampers the highly accurate evaluation of automatic algorithms and, as shown through comparison with our GTA-2 annotations and in several clinical studies, it is suboptimal with respect to the accuracy achievable by a human observer.

On the other hand, the great majority of annotations in GTA-1 are approximately correct and only in rare cases would we find annotations clearly off-target. This, together with the impractical task of re-annotating all FRGC scans (nearly 5000) motivates interest in trying to derive maximum benefit from the available annotations. In the next section we present an algorithm to reduce the negative effects of annotation errors for learning-based models for automatic landmark detection.

III. Training local descriptors with uncertain annotations

Let $\{\mathscr{M}_i\}_{i=1}^N$ be a set of facial surfaces described by vertices $\mathbf{v} \in \mathscr{M}_i$, let $\{\mathbf{a}_i\}_{i=1}^N$ be the set of corresponding 3D annotations containing L landmarks each and $D(\mathbf{v})$ be a *descriptor* that can be computed for every vertex $\mathbf{v}$. For example, spin images [10] or 3D shape contexts [5] are some popular geometric descriptors.

We wish to train a local descriptor model for each landmark. The objective is being able to compute a *similarity score* $s(\mathbf{v})$ based solely on the local descriptors that correlates

[1] Annotations available at http://fsukno.atspace.eu/Data.htm

[2] Available at http://www-users.cs.york.ac.uk/∼creusot/

[3] Given 100 scans there are $\binom{100}{2} = 4950$ distinct pairs. When considering the first 100 scans of FRGC, 248 of those pairs are composed by scans of the same person.

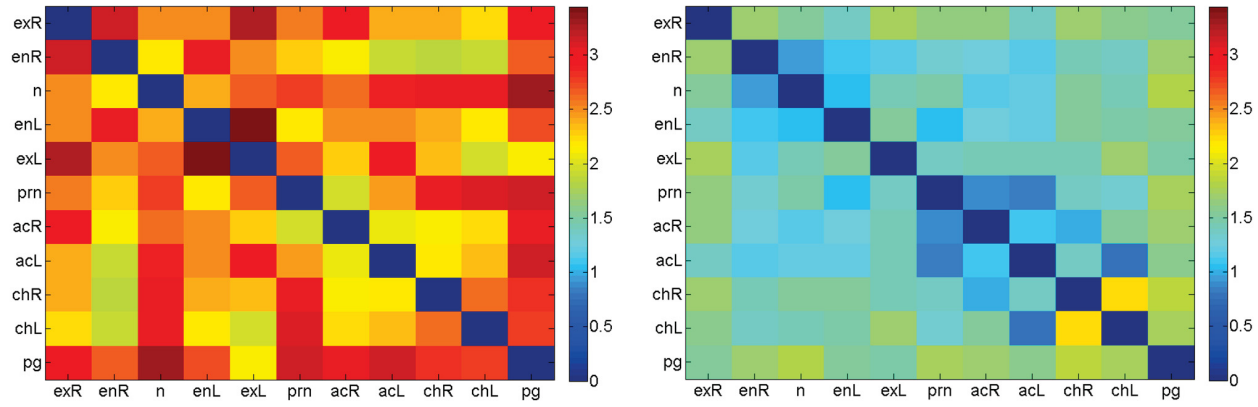

Fig. 1. Inter-landmark distance errors averaged over all possible pairs of replicates for the first 100 scans from FRGC dataset using public annotations (GTA-1, left) and our own manual annotations (GTA-2, right). Errors are color coded according to the legend, in mm. The following 11 landmarks are evaluated: outer eye corners or *exocanthion* (ex, Left & Right), inner eye corners or *endocanthion* (en, Left & Right), nose root or *nasion* (n), nose tip or *pronasale* (prn), nose corners or *alare crest* (ac, Left & Right), mouth corners or *cheilion* (ch, Left & Right) and chin tip or *pogonion* (pg).

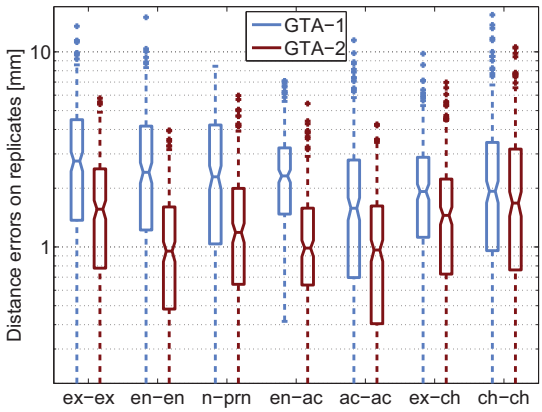

Fig. 2. Boxplots of inter-landmark distance errors on the 248 pairwise comparison between replicates that can be computed from the first 100 scans from FRGC.

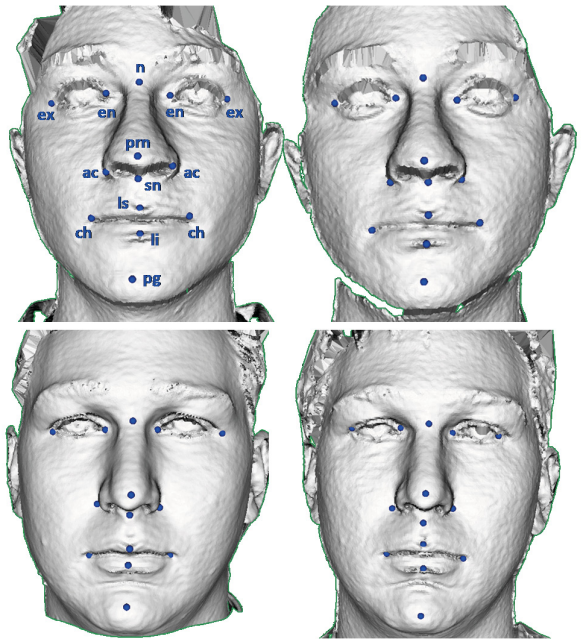

Fig. 3. Two examples of replicates with non-consistent annotations. Each row shows two scans from the same person selected to highlight different criteria when targeting the same landmarks. Note especially nose corners (ac), superior and inferior lips (ls, li) in the top row; eye corners (en, ex), subnasale (sn) and chin tip (pg) in the bottom row. Landmark abbreviations are indicated in the top-left scan.

well with the distance to the correct position of the targeted landmark. That is, for each landmark ℓ_p we seek for a function $f_p()$ so that $s(\mathbf{v}) = f_p\big(D(\mathbf{v})\big)$ is high for vertices close to $\mathbf{a}(\ell_p)$ and low for all other vertices of the mesh.

One of the simplest options, quite widespread both in 2D and 3D landmark localization literature, is to compute the similarity scores as the distance to a *template* derived as the average descriptor from the manual annotations. That is:

$$f_p\big(D(\mathbf{v}), \overline{\mathbf{D}}_p\big) = \|D(\mathbf{v}) - \overline{\mathbf{D}}_p\| \tag{4}$$

$$\overline{\mathbf{D}}_p = \frac{1}{N}\sum_{i=1}^{N} D(\mathbf{a}_i(\ell_p)) \tag{5}$$

In the above expression we fully trust the correctness of manual annotations. As we saw in the previous section, this might not be the case and we need to account for uncertainty in the ground truth. We do so by assuming that the *correct landmark position* does not necessarily coincide with annotations in $\{\mathbf{a}_i\}$ but are within an *uncertainty radius* r_u from them. Let $\{\tilde{\mathbf{a}}_i\}$ be those hypothetically correct (but unknown) landmark positions. Then we assume that:

$$d\big(\mathbf{a}_i(\ell_p), \tilde{\mathbf{a}}_i(\ell_p)\big) \leq r_u \tag{6}$$

and build an average descriptor template as in eq. (5) but allowing the position of landamrks to be updated to any vertex within r_u from the original annotations. Evidently the key point is to choose the appropriate vertices, for which we assume

that the majority of annotations are approximately correct and distributed around the target (hypothetically correct) position without significant bias (as otherwise the biased position would become the hypothetical target). We determine the best vertices to use for computation of the template by minimizing the annotation corrections that their choice would implicitly suggest, as detailed below.

Let us hypothesize that the *j-th* scan is correctly annotated for landmark ℓ_p and we choose its descriptor as $\overline{\mathbf{D}}_p = D(\mathbf{a}_j(\ell_p))$. Assuming that the local descriptor is good enough to provide an acceptable estimate of corresponding points between meshes, we can estimate a new set of annotations, the *j-th corrected set* $\{\hat{\mathbf{a}}_i^j(\ell_p)\}$, as the vertices within r_u that maximize the similarity score $s(\mathbf{v})$:

$$\hat{\mathbf{a}}_i^j(\ell_p) = \arg\max_{\mathbf{v}\in\mathscr{N}_p} f\big(D(\mathbf{v}), \mathbf{a}_j(\ell_p)\big)\big) \tag{7}$$

$$\mathscr{N}_p = \big\{\mathbf{v}\in\mathscr{M}_i \,|\, d\big(\mathbf{v}, \mathbf{a}_i(\ell_p)\big) \le r_u\big\} \tag{8}$$

That is, from the point of view of the template chosen to compute the similarity scores, the annotations should be *corrected* as indicated above, as we are implicitly assuming that $\{\hat{\mathbf{a}}_i\}$ would be an estimate of $\{\tilde{\mathbf{a}}_i\}$. Since we assumed that the majority of annotations are approximately correct, it is sensible to choose the *j-th* scan as the one whose annotations minimize some measure of the induced corrections, such as:

$$j_p = \arg\min_j \frac{1}{N}\sum_{i=1}^{N}\big(d\big(\hat{\mathbf{a}}_i^j(\ell_p) - \mathbf{a}_i(\ell_p)\big)\big)^2 \tag{9}$$

Thus, we end up with a new set of annotations obtained as the Least Squared Corrections of Uncertainty (LSCU). We can now estimate the template to use for the *p-th* landmark as the average descriptor from the set of corrected annotations induced by the j_p*-th* scan:

$$\overline{\mathbf{D}}_p = \frac{1}{N}\sum_{i=1}^{N} D(\hat{\mathbf{a}}_i^{j_p}(\ell_p)) \tag{10}$$

We found experimentally that both for the minimization in eq. (9) and for obtaining the descriptor template in eq. (10) it is advantageous to use the median instead of the mean (probably due to its robustness to outliers). Hence, all results reported in this paper are based on the median.

IV. Experimental results

We tested the LSCU algorithm introduced in Section III in the context of automatic landmark localization using the SRILF algorithm, which is briefly described in the next subsection for completeness.

A. Shape Regression with Incomplete Local Features (SRILF)

The SRILF algorithm [20] combines the response from local feature detectors for each of the targeted landmarks with statistical constraints that ensure the plausibility of landmark positions on a global basis. The algorithm has three components: *i*) selection of candidates through local feature detection; *ii*) partial set matching to infer possibly missing landmarks; *iii*) combinatorial search, which integrates the other two components.

The selection of candidates is performed independently for each targeted landmark; a similarity score is computed for every vertex and the top-scoring ones are retained as candidates for the considered landmark. As in many other algorithms, it is expected that one of these candidates will be close enough to the correct position of the landmark. Nonetheless, the number of false positives (i.e. vertices that produce high similarity scores even though they are far from the correct landmark location) can change considerably for different landmarks, as well as from one facial scan to another, making it difficult to choose the number of candidates that should be retained.

While many approaches try to retain large numbers of candidates to make sure that at least one will be *reasonably close* to the desired landmark position, SRILF determines the number of candidates as an upper outlier threshold from the distribution of false positives over a training set. This implies that, in the vast majority of cases, a candidate that is close enough to the target landmark will be detected, but a small proportion will be missed. Hence, for each targeted landmark there will be an initial set of candidates that may or may not contain a suitable solution and we need to match our set of target landmarks to a set of candidates that is potentially incomplete. This is analogous to the point-matching problem found in algorithms that search for correspondences. However, the human face is a non-rigid object and these point-matching algorithms are typically restricted to rigid transformations.

The second component of the algorithm aims at dealing with the above problem. Based on the priors encoded in a statistical shape model, it uses a subset of the landmarks (i.e. those with suitable candidates) to infer the most likely position of the ones that are missing.

Finally, the third component of the algorithm integrates the two previous steps into a combinatorial search. It consists of analyzing subsets of candidates and completing the missing information by inferring the coordinates that maximize the probability of a deformable shape model. Thus, despite the resulting subset possibly containing only part of the targeted landmarks, estimates for the remaining coordinates are inferred by regression from the priors encoded in the model. Subsets of candidates that fulfill the statistical constraints of the model are retained and additional landmarks are incorporated iteratively as long as the set remains a plausible instance of the shape model, in a sequential forward selection strategy. The cost of including a new candidate is computed as the median of squared distances to the closest candidate (per landmark), which provides robustness to potential outliers (e.g. landmarks for which no nearby candidates have been found). The best solution is the one with minimum inclusion cost among those with the largest number of candidates (i.e. those with the largest support).

B. Effect of uncertainty handling on localization results

Our first set of experiments aims to evaluate the impact of the LSCU correction algorithm presented in Section III on the localization accuracy that can be obtained. To do this, we trained the SRILF algorithm using the corrected annotations $\{\hat{\mathbf{a}}_i\}$ for different values of the uncertainty radius r_u. Spin images were used as the local descriptors, following the settings in [20] and using cross-correlation to the corresponding template to compute the similarity scores, as this is

the metric originally proposed for spin images [10]. We should emphasize that, although results would change depending on the descriptor, the correction method itself is not restricted to this particular choice.

Experiments were carried out using the first 100 scans from FRGC (as in Section II), performing a 6-fold cross-validation so that no scan is included in training and test sets at the same time. All scans were pre-processed with a median filter to remove spikes (as suggested in [19]) and a smoothing filter based on a bi-quadric approximation of each vertex from a 3 mm neighborhood. Finally, scans were decimated by a factor of 4 : 1 and converted to triangulated meshes. This resulted in an average of approximately 22000 vertices per mesh.

Localization accuracy was measured as the point-to-point Euclidean distance between the landmarks automatically located by SRILF and the original set of manual annotations $\{\mathbf{a}_i\}$, i.e. corrections of the annotation set are used exclusively for training and never for testing.

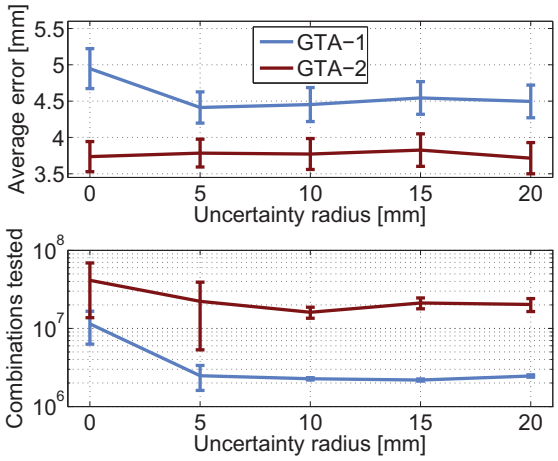

Fig. 4. Localization errors (mean with 95% confidence intervals) using SRILF averaged over all landmarks (top) and number of combinations tested (bottom) for different values of the r_u (uncertainty radius). When $r_u = 0$ the original annotations are trusted, while for $r_u > 0$ the SRILF algorithm is trained with the annotations corrected by LSCU. Errors are always computed with respect to the original set of annotations.

Fig. 4 shows the localization results averaged for all landmarks using different values of r_u up to 20 mm. Notice that when $r_u = 0$ we do not allow any correction of the training set and rely completely on the original annotations. Several conclusions can be extracted from this plot. Firstly, the introduction of corrections in the training set allows for a significant reduction in average error when using the *noisy* annotations in GTA-1 but produces little or no change in our annotated set, GTA-2. In both cases the behavior is relatively stable for a wide range of uncertainty radii (at least up to 20 mm), which is very important for the practical applicability of LSCU as r_u is the only parameter to be set.

It is also clear that models built using the more consistent annotations in GTA-2 obtain higher accuracy than models built from annotations in GTA-1. This was expected from the results presented in Section II. Let us emphasize that this does not necessarily imply that annotations from GTA-2 are more accurate than GTA-1, but they are more precise and would therefore be easier to learn for an automatic algorithm.

Fig. 4 also includes the average number of combinations that were tested for each setting. In the SRILF algorithm. The number of combinations to test relates to the number of false positives in the local descriptors and is a measure of the computational complexity associated with the constructed model [20]. Our results show that correction of the training sets also allows reducing the number of tested combinations and therefore the computational complexity.

Separate results for the different landmarks are provided in Table I with results for the symmetric landmarks (i.e. left & right) merged together by averaging. When using annotations from GTA-1, the correction of the training set helps reducing the errors in all landmarks but the chin tip, which hardly changes. In several cases errors were reduced more than 10% and in a majority of them differences were statistically significant[4].

Table I also includes the results obtained using GTA-2. As suggested by the plots in Fig. 4, localization results did not vary much when introducing corrections to GTA-2 so we only include the case of $r_u = 0$. In general, errors obtained using GTA-2 were more than 20% lower than those obtained using GTA-1. Using LSCU to correct the annotations helped to reduce this gap but only by approximately half.

Nonetheless, we should recall that errors from GTA-1 and GTA-2 are not directly comparable since there is considerable difference in the consistency of both sets of annotations. To investigate whether automatic corrections would converge to the accuracy achievable with the *cleaner* annotations from GTA-2, we repeated the computation of errors using always the annotations from GTA-2 as ground truth, regardless of having used GTA-1 or GTA-2 for training. The results are displayed in Fig. 5 and Table II. It can be observed that now the overall errors from the models trained on GTA-1 with automatic corrections tend to converge to the errors of models trained on GTA-2. We can see in Table II that this is also true for 8 out of the 11 tested landmarks, which is striking. Only the outer eye corners (*ex*) and the chin tip (*pg*) do not reach the accuracy obtained with annotations from GTA-2 (indeed the latter one shows some increase of the error, although not statistically significant). It is interesting to note that *ex* and *pg* have been shown to be among the most difficult points to locate using spin images [21], which may explain these results.

C. Comparison to Support Vector Machines

A potential weakness of the method proposed in Section III is the rather simplistic formulation used to derive the template $\overline{\mathbf{D}}_p$ as the average (or the median) of descriptors from the training set. Thus, although the results reported so far demonstrate the effectiveness of LSCU to correct the annotations, it might be worth investigating whether more powerful machine learning techniques can directly resolve the lack of consistency of GTA-1 and train highly accurate models.

To test the above hypothesis we replaced computation of the similarity scores described in Section III by a classifier based on Support Vector Machines (SVM) [3]. It should be

[4] $p < 0.05$, paired Wilcoxon signed rank test.

TABLE I. LANDMARK LOCALIZATION ERRORS FOR DIFFERENT VALUES OF THE UNCERTAINTY RADIUS. FOR EACH CELL, THE TOP ROW INDICATES THE AVERAGE ± STANDARD ERROR [MM] AND THE BOTTOM ROW SHOWS THE ERROR CHANGE WITH RESPECT TO THE RESULTS USING GTA-1 WITHOUT CORRECTION, WHICH ARE TAKEN AS REFERENCE. ASTERISKS INDICATE STATISTICALLY SIGNIFICANT DIFFERENCES AT $p < 0.05$

Landmark	GTA-1					GTA-2
	$r_u = 0$	$r_u = 5mm$	$r_u = 10mm$	$r_u = 15mm$	$r_u = 20mm$	$r_u = 0$
ex	6.25 ± 0.29 (ref)	5.56 ± 0.26 −11.0% *	5.69 ± 0.28 −8.9% *	5.78 ± 0.28 −7.4% *	5.82 ± 0.28 −6.8%	4.85 ± 0.21 −22.4% *
en	4.88 ± 0.20 (ref)	4.64 ± 0.19 −4.9%	4.44 ± 0.18 −9.0% *	4.41 ± 0.17 −9.5% *	4.65 ± 0.19 −4.6%	3.78 ± 0.18 −22.4% *
n	4.19 ± 0.22 (ref)	3.35 ± 0.18 −20.2% *	3.70 ± 0.19 −11.8% *	3.59 ± 0.18 −14.3% *	3.66 ± 0.19 −12.8% *	3.09 ± 0.16 −26.3% *
prn	4.02 ± 0.31 (ref)	3.63 ± 0.21 −9.7%	3.49 ± 0.20 −13.2%	3.58 ± 0.19 −10.9%	3.44 ± 0.18 −14.5%	3.10 ± 0.19 −23.0% *
ac	4.05 ± 0.21 (ref)	3.63 ± 0.24 −10.4% *	3.66 ± 0.22 −9.6% *	3.70 ± 0.22 −8.6%	3.52 ± 0.21 −13.1% *	3.12 ± 0.16 −22.9% *
ch	5.32 ± 0.31 (ref)	4.41 ± 0.23 −17.0% *	4.39 ± 0.21 −17.4% *	4.75 ± 0.24 −10.7% *	4.47 ± 0.24 −16.0% *	3.64 ± 0.20 −31.6% *
pg	5.23 ± 0.35 (ref)	5.09 ± 0.33 −2.7%	5.43 ± 0.33 +3.8%	5.51 ± 0.34 +5.4%	5.44 ± 0.35 +4.0%	4.12 ± 0.25 −21.2% *

TABLE II. LANDMARK LOCALIZATION ERRORS FOR DIFFERENT VALUES OF THE UNCERTAINTY RADIUS. FOR EACH CELL, THE TOP ROW INDICATES THE AVERAGE ± STANDARD ERROR [MM] AND THE BOTTOM ROW SHOWS THE ERROR CHANGE WITH RESPECT TO THE RESULTS USING GTA-1 WITHOUT CORRECTION, WHICH ARE TAKEN AS REFERENCE. ASTERISKS INDICATE STATISTICALLY SIGNIFICANT DIFFERENCES AT $p < 0.05$

Landmark	Train with GTA-1, test on GTA-2					GTA-2
	$r_u = 0$	$r_u = 5mm$	$r_u = 10mm$	$r_u = 15mm$	$r_u = 20mm$	$r_u = 0$
ex	5.97 ± 0.29 (ref)	5.69 ± 0.23 −4.6%	5.59 ± 0.24 −6.3%	5.30 ± 0.23 −11.2% *	5.33 ± 0.24 −10.7% *	4.85 ± 0.21 −18.8% *
en	4.47 ± 0.15 (ref)	4.17 ± 0.17 −6.9%	3.88 ± 0.14 −13.4% *	3.79 ± 0.13 −15.4% *	3.92 ± 0.14 −12.3% *	3.78 ± 0.18 −15.4% *
n	3.97 ± 0.22 (ref)	2.96 ± 0.17 −25.5% *	2.98 ± 0.20 −25.1% *	2.86 ± 0.20 −27.9% *	2.97 ± 0.18 −25.3% *	3.09 ± 0.16 −22.3% *
prn	3.47 ± 0.28 (ref)	3.16 ± 0.21 −9.0%	2.97 ± 0.19 −14.4% *	3.06 ± 0.18 −11.8%	2.97 ± 0.17 −14.5% *	3.10 ± 0.19 −10.7% *
ac	3.41 ± 0.15 (ref)	2.93 ± 0.11 −14.1% *	3.04 ± 0.12 −10.9% *	2.98 ± 0.12 −12.4% *	2.92 ± 0.12 −14.3% *	3.12 ± 0.16 −8.3% *
ch	4.45 ± 0.29 (ref)	3.59 ± 0.22 −19.4% *	3.59 ± 0.18 −19.3% *	3.83 ± 0.19 −14.0% *	3.48 ± 0.20 −21.8% *	3.64 ± 0.20 −18.2% *
pg	4.87 ± 0.33 (ref)	5.03 ± 0.31 +3.4%	5.21 ± 0.29 +7.0%	5.35 ± 0.30 +10.0%	5.05 ± 0.28 +3.8%	4.12 ± 0.25 −15.3% *

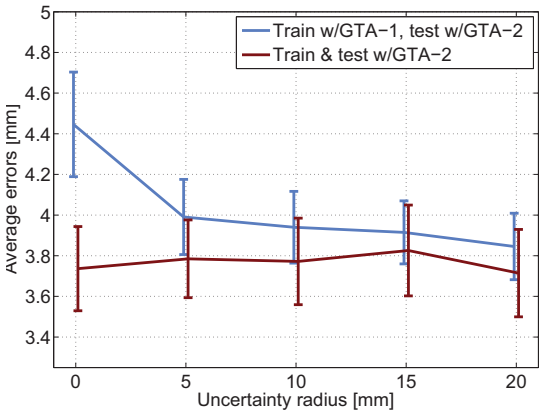

Fig. 5. Localization errors (mean with 95% confidence intervals) using SRILF averaged over all landmarks for different values of the r_u (uncertainty radius). When $r_u = 0$ the original annotations are trusted, while for $r_u > 0$ the SRILF algorithm is trained with the annotations corrected by LSCU. Errors are always computed with respect to the original annotations from GTA-2, regardless of the training set that is used.

noted that the standard solution of training a 2-class SVM is not appropriate for our problem since it would require a threshold to separate vertices between positive and negative examples and to employ decimation techniques to reduce the size (and the unbalance) of the resulting training sets[5].

Thus, we chose to build a regression model rather than a discrete classifier. The training set was composed of all vertices within a neighborhood of 5 mm from the ground truth annotations, using the distances to the ground truth as labels to train the regressor. Specifically, we used the ν-Support Vectors Regression [17] as implemented in the publicly available SVM library [2] with the default kernel (Radial Basis Functions) and $\nu = 0.02$, as it was experimentally determined that this value would limit the number of support vectors to between 100 and 200, which is comparable to the number of scans in the training set and should avoid over-fitting. Notice that, for these tests, we do not attempt any correction of the annotated landmarks (i.e. $r_u = 0$).

Table III shows the results for both sets of ground truth annotations, as well as results when training on GTA-1 and testing on GTA-2. Compared to the errors reported in the previous section, we can see that:

- The use of SVMs results in an increase of localization accuracy both on the GTA-1 and GTA-2 datasets.
- The errors obtained when training with GTA-2 are still much lower than those obtained when training with GTA-1. In relative terms, reduction in the error is

[5]Recall that a typical facial scan has approximately 22000 vertices and, theoretically, only one of them (two in the case of symmetric landmarks) would be a positive example.

TABLE III. LANDMARK LOCALIZATION ERRORS USING SVM REGRESSION TO COMPUTE THE DESCRIPTOR SIMILARITY SCORES. AVERAGES ± STANDARD ERRORS [MM] ARE SHOWN FOR EACH LANDMARK AND ALSO AVERAGED OVER ALL 11 LANDMARKS.

Train with	GTA-1		GTA-2
Test with	GTA-1	GTA-2	
ex	5.36 ± 0.21	6.41 ± 0.25	4.20 ± 0.22
en	4.10 ± 0.17	4.12 ± 0.17	3.01 ± 0.17
n	2.91 ± 0.20	2.32 ± 0.17	2.17 ± 0.22
prn	3.36 ± 0.21	3.34 ± 0.18	2.46 ± 0.22
ac	3.50 ± 0.23	2.93 ± 0.15	2.93 ± 0.18
ch	4.72 ± 0.31	4.28 ± 0.29	2.78 ± 0.21
pg	5.38 ± 0.39	4.83 ± 0.37	3.79 ± 0.22
Overall	4.28 ± 0.14	4.18 ± 0.13	3.12 ± 0.12

very similar to that observed when using the average template.

Hence, the use of SVMs improves accuracy but suffers from the inconsistency of annotations in a similar proportion as for the correlation to the average template.

It is interesting to compare the results obtained by models trained with GTA-1 when tested on the *noisy* GTA-1 or on the *cleaner* GTA-2. When testing on GTA-1, SVMs perform slightly better than average template models even after correction. However, when testing on GTA-2, average template models with corrections clearly outperform SVMs. This suggests that SVM classifiers might be partially learning the noise in the annotations (at the expense of additional complexity), while LSCU is actually simplifying these based on the consistency of the local descriptors.

It is worth mentioning that the complexity involved in the use of SVMs was some orders of magnitude above the complexity using the averaged templates. The use of averaged templates requires the computation of a distance metric (correlation in our case) with respect to a single reference (the template), while the resulting SVMs in our experiments averaged 223 and 161 support vectors for models trained on GTA-1 and GTA-2, respectively. Although computation time is not the main focus of this work, we note that evaluating all vertices of the surfaces like those used in our experiments with such SVM models can take of the order of one minute on a standard computer, while it takes less than a second to evaluate these based on an average template.

V. CONCLUSIONS

We present an algorithm aimed at correcting a set of manual annotations, with the goal of enhancing accuracy in the automatic models built from them. Experiments on a set of *noisy* annotations publicly available for 100 scans in the FRGC database show that models built from annotations corrected by LSCU are significantly more accurate than models built from the original annotations. The only parameter of the algorithm, the uncertainty radius, controls the maximum displacement that is allowed for the corrections and we show that its choice has a fairly limited impact.

Results from the public annotations are compared to our own set of manual annotations. We objectively show that the latter has higher consistency, which allows construction of more accurate models. Applying LSCU to this set of *cleaner* annotations did not produce significant changes, which suggests that the algorithm does not distort the input data. Additionally, we showed that by applying LSCU to the public annotations, it is possible to build models that obtain accuracy similar to those built on our own set of cleaner annotations. Additionally, as indicated in Section II, we make this new set of annotations available.

The above conclusions apply to overall performance across a set of 11 prominent facial landmarks and also individually to 8 of these: inner eye corners, mouth corners, nose tip and corners and nose root. Regarding the other 3 landmarks, outer eye corners were also improved but to a lesser extent (i.e. they did not converge to the performance achieved with the cleaner set) and the chin tip showed slight impairment. Such results correlate well with the lower performance of spin images for these 3 landmarks and suggest that here we might not be able to produce acceptable correspondences across meshes, as was assumed in the derivation of the LSCU algorithm.

We emphasize that our evaluation is based purely on the effects of LSCU as a pre-processing step to increase the accuracy of automatic models (namely, the local geometry descriptors used by it) and does not imply that the resulting annotations could be used to replace the ground truth for evaluation purposes. The objective of LSCU is to enhance consistency of the local descriptors extracted from the annotations, hence it does not guarantee correctness of the resulting landmarks. Indeed, we have chosen rather simple components to test our algorithm, such as spin images and an average template. Thus, it should be clear that the resulting corrections cannot serve as alternative ground truth. Nonetheless, the excellent results obtained in terms of model training encourage further research and suggest that more elaborate models might produce corrected sets that are able to challenge manual annotations in a wider context.

ACKNOWLEDGMENTS

The authors would like to thank their colleagues in the Face3D Consortium (www.face3d.ac.uk), and the financial support provided from the Wellcome Trust (grant 086901/Z/08/Z) and the Marie Curie IEF programme (grant 299605, SP-MORPH).

REFERENCES

[1] C.E. Brodley and M.A. Friendl. Comparing boosting and bagging techniques with noisy and imbalanced data. *Journal of Artificial Intelligence Research*, 11:131–167, 1999.

[2] C.C. Chang and C.J Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011.

[3] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20(3):273–297, 1995.

[4] C. Creusot, N. Pears, and J. Austin. Automatic keypoint detection on 3D faces using a dictionary of local shapes. In *Proc. 1st Joint Conf. on 3D Imaging, Modeling, Processing, Visualization, and Transmission, Hangzhou, China*, pages 204–211, 2011.

[5] A. Frome, D. Huber, R. Kolluri, T. Bulow, and J. Malik. Recognizing objects in range data using regional point descriptors. In *Proc. 8th European Conf. on Computer Vision, Prague, Czech Republic. LNCS vol. 3023*, pages 224–237, 2004.

[6] D. Gamberger, N. Lavrac, and S. Dzeroski. Noise detection and elimination in data preprocessing: experiments in medical domains. *Applied Artificial Intelligence*, 14:205–223, 2000.

[7] S. Gupta, M.K. Markey, and A.C. Bovik. Antopometric 3D face recognition. *International Journal of Computer Vision*, 90(3):331–349, 2010.

[8] C.L. Heike, M.L. Cunningham, A.V. Hing, E. Stuhaug, and J. Starr. Picture perfect? reliability of craniofacial anthropometry using three-dimensional digital stereophotogrammetry. *Plastic & Reconstructive Surgery*, 124(4):1261–1272, 2009.

[9] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano. Skewed class distributions and mislabeled examples. In *Proc. 7th IEEE Int. Conf. on Data Mining Workshops, Omaha, NE, USA*, pages 477–482, 2007.

[10] A.E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):433–449, 1999.

[11] T.M. Khoshogoftaar, J. van Hulse, and A. Napolitano. Comparing boosting and bagging techniques with noisy and imbalanced data. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, 41(3):552–568, 2011.

[12] A. Malossini, E. Blanzieri, and R.T. Ng. Detecting potential labeling errors in microarrays by data perturbation. *Bioinformatics*, 22(17):2114–2121, 2006.

[13] D.F. Nettleton, A. Orriols-Puig, and A. Fornells. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial Intelligence Review*, 33(4):275–306, 2010.

[14] G. Passalis, N. Perakis, T. Theoharis, and I.A. Kakadiaris. Using facial symmetry to handle pose variations in real-world 3D face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1938–1951, 2011.

[15] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowver, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recogmition grand challenge. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, San Diego, CA, USA*, volume 1, pages 947–954, 2005.

[16] J.M. Plooij, G.R.J. Swennen, F.A. Rangel, T.J.J. Maal, F.A.C. Schutyser, E.M. Bronkhorst, A.M. Kuijpers-Jagtman, and S.J. Bergé. Evaluation of reproducibility and reliability of 3D soft tissue analysis using 3D stereophotogrammetry. *International Journal of Oral and Maxillofacial Surgery*, 38(3):267–273, 2009.

[17] B. Scholkopf, A. Smola, R.C. Williamson, and P.L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

[18] N. Segata, E. Blanzieri, S.J. Delany, and P. Cunningham. Noise reduction for insance-based learning with a local maximal margin approach. *Journal of Intelligent Information Systems*, 35:301–331, 2010.

[19] M.P. Segundo, L. Silva, O.R. Pereira Bellon, and C.C. Queirolo. Automatic face segmentation and facial landmark detection in range images. *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics*, 40(5):1319–1330, 2010.

[20] F.M. Sukno, J.L Waddington, and P.F. Whelan. 3D facial landmark localization using combinatorial search and shape regression. In *Proc. ECCV Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment, Firenze, Italy. LNCS vol. 7583*, pages 32–41, 2012.

[21] F.M. Sukno, J.L Waddington, and P.F. Whelan. Comparing 3D descriptors for local search of craniofacial landmarks. In *Proc. 8th Int. Symp. on Visual Computing, Rethymnon, Crete, Greece. LNCS vol. 7432*, pages 92–103, 2012.

[22] P. Szeptycki, M. Ardabilian, and L. Chen. A coarse-to-fine curvature analysis-based rotation invariant 3D face landmarking. In *Proc. 3rd IEEE Int. Conf. on Biometrics: Theory, Applications and Systems, Washington DC, USA*, pages 1–6, 2009.

[23] A.M. Toma, A. Zhurov, R. Playle, E. Ong, and S. Richmond. Reproducibility of facial soft tissue landmarks on 3D lasser-scanned facial images. *Orthodontics & Craniofacial Research*, 12(1):33–42, 2009.

[24] S. Verbaeten and A. van Assche. Ensemble methods for noise elimination in classification problems. In *Proc. 4th Int. Workshop on Multiple Classifier Systems, Guildford, UK. LNCS vol. 2709*, pages 317–325, 2003.

[25] T.H. Yu and Y.S. Moon. A novel genetic algorithm for 3D facial landmark localization. In *Proc. 2nd IEEE Int. Conf. on Biometrics: Theory, Applications and Systems, Arlington, VA, USA*, pages 1–6, 2008.

[26] X. Zhao, P. Szeptycki, E. Dellandrea, and L. Chen. Precise 2.5D facial landmarking via an analysis by synthesis approach. In *Proc. Workshop on Applications of Computer Vision, Sonowbird, UT, USA*, pages 1–7, 2009.